Darwintrees for action recognition

Albert Clapés[1,2], Tinne Tuytelaars[3], Sergio Escalera[1,2]

[1]*Dept. of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain* | [2]*Comuter Vision Center, Autonomous University of Barcelona, Bellaterra (Barcelona), Spain* | [3]*Katholieke Universiteit Leuven, ESAT department-PSI, imec, Leuven, Belgium*

International Conference on Computer Vision 2017

Section 1: Introduction

- **Proposal**: a novel **mid-level representation for action/activity recognition** on RGB videos on the basis of *improved dense trajectories* (IDT) [1], *fisher vectors* (FV), and *videodarwin* (VD) [2].
- We model the evolution of features not only for the entire video, but also on its subparts (represented as nodes in a binary tree hierarchically grouping subsets of IDTs).
- For each node, we compute Node-VD and Branch-VD. These are later combined with with VD on the whole video trajectories (Root-VD) a to perform classification with SVM.
- Results: better performance than standard VD (i.e., global-VD) and defines the state-of-the-art on *UCF-Sports* [3] and *Highfive* [4] action recognition datasets.

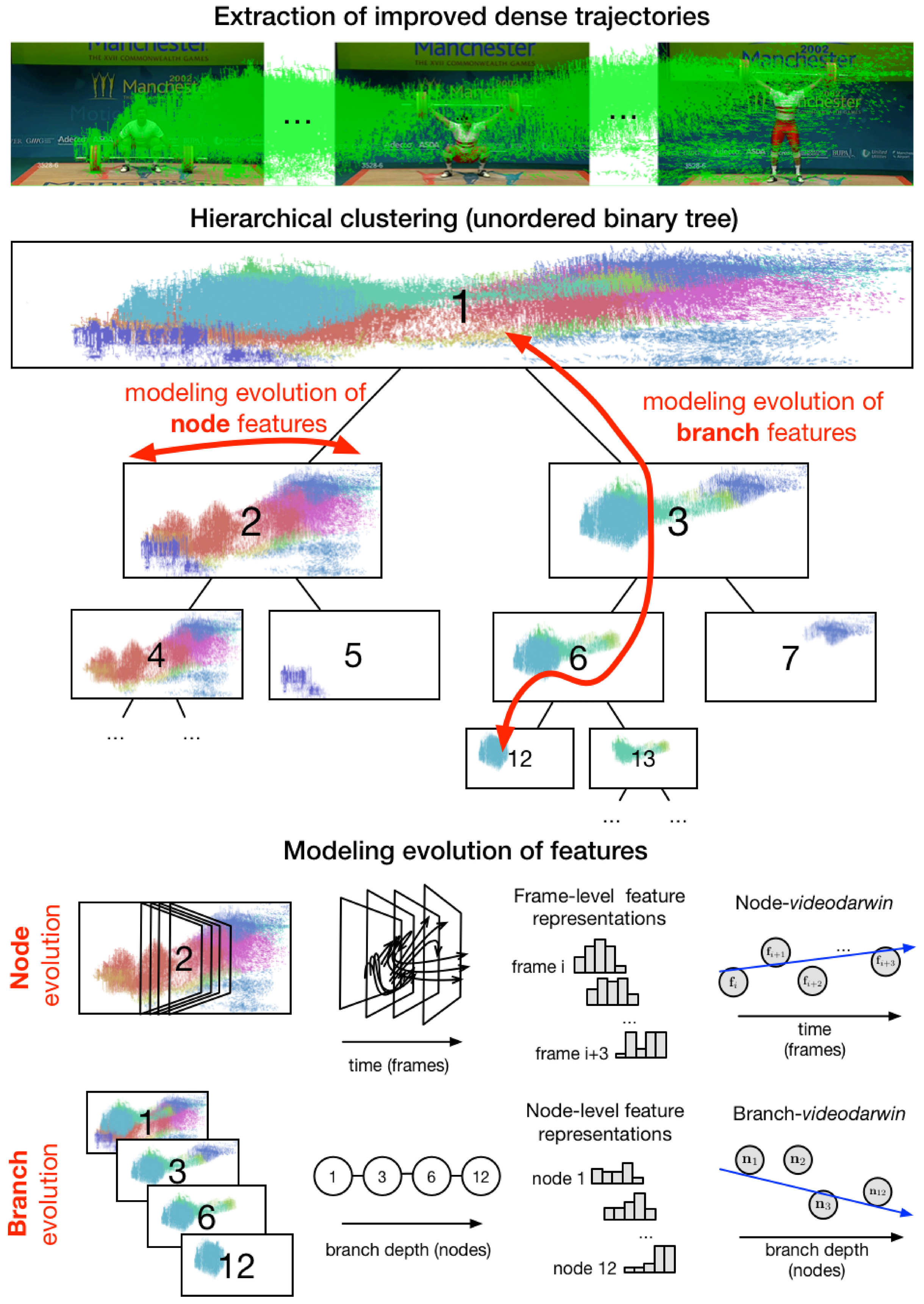

Fig. 1. The pipeline. Each leaf node is represented in a different color.

Section 2: Method

Binary tree of trajectory construction

- By recursively applying a *divisive spectral clustering algorithm* [5] on the set of trajectories D.
- For the clustering, we used primitive trajectory features $\bar{x}, \bar{y}, \bar{t}, \bar{v}_x, \bar{v}_y$.
- A tree node i containing the set of trajectories $D_i \subseteq D$ expands a temporal segment (t_i, t'_i) of the T-frame video, $0 \leq t_i < t'_i < T_i$.
- Let $\mathbf{U}_i$ and $\mathbf{u}_i$ be respectively the matrix of **per-frame FVs** and the **global FV** on D_i.

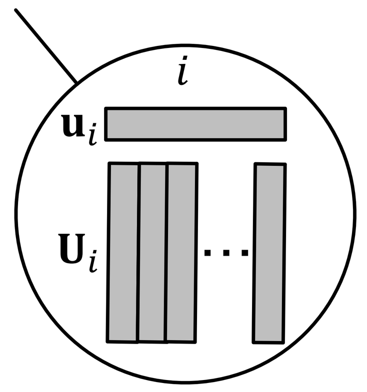

Fig. 2. i-th node representation: global FV for all IDTs assigned to the node's cluster, $\mathbf{u}_i$, and matrix of per-frame FVs, $\mathbf{U}_i$.

Videodarwin: in-a-nutshell

- VD applies any learning algorithm able to model frame ordering in a sequence. Our choice is to use a *linear regressor* we refer to as ν.
- We compute VD in forward and reverse directions.
- Prior to VD, *time varying mean* is applied. Given $\mathbf{X} \epsilon \mathbb{R}^{\#\{\text{features}\} \times \#\{\text{timesteps}\}}$, forward videodarwin (FW) is calculated as follows:

$$\mathbf{m}_\tau^{FW} = \frac{1}{\tau} \sum_{k=1}^{\tau} \mathbf{X}_{:,k}$$

$$\mathbf{V}_{:,\tau}^{FW} = \frac{\mathbf{m}_\tau}{||\mathbf{m}_\tau||_1}, \forall \tau = 1, \ldots, T$$

Note reverse VD simply re-defines m_τ^{FW} to calculate the varying mean backwards.

- The final VD representation, $\mathbf{w}$, is then:

$$\mathbf{w}^{FW} = \nu(\mathbf{V}^{FW}, (1..T))$$

$$\mathbf{w}^{RV} = \nu(\mathbf{V}^{RV}, (1..T))$$

$$\mathbf{w} = [\mathbf{w}^{FW}; \mathbf{w}^{RV}]$$

Mid-level representations

- **Node-VD** representation on node i, i.e. $\mathbf{n}_i$, by taking $\mathbf{X} = \mathbf{U}_i$. In particular, **Root-VD** is just the special case $i = 1$.
- **Branch-VD** on node i requires its ancestors to be represented by their global FV, $\mathbf{u}_i$. We construct i-th node's branch as a matrix of per-node global FVs. That is:

$$\mathbf{B}_i = [\mathbf{u}_i, \mathbf{u}_{i//2^1}, \mathbf{u}_{i//2^2}, \ldots, \mathbf{u}_1]$$

- Then, i-th node's branch representation, $\mathbf{b}_i$, is computed taking $\mathbf{X} = \mathbf{B}_i$.

Darwintree kernel classification

- Each tree has an arbitrary number of nodes and each node is represented by the combination of Node- and Branch-VD:

$$\mathbf{s}_i = [\mathbf{n}_i; \mathbf{b}_i], i > 1.$$

- We define the ***Darwintree kernel*** function k_{DT} between two trees (S, S') based on pairwise similarities of their nodes' representations:

$$k_{\mathrm{DT}}(S, S') = \frac{1}{|S||S'|} \sum_{\mathbf{s}_i \epsilon S} \sum_{\mathbf{s}_j \epsilon S'} \phi(\mathbf{s}_i, \mathbf{s}_j),$$

$\forall i, j > 1$, where $\phi(\cdot,\cdot)$ can be any linear mapping function (e.g. dot product).

Since root node has no ancestors, we define a different kernel:

$$k_{\mathrm{root}}(\mathbf{n}_1, \mathbf{n}'_1) = \phi(\mathbf{n}_1, \mathbf{n}'_1)$$

- Finally, a **linear SVM** performs classification using a linear combination of k_{DT} and k_{root}:

$$k_{\mathrm{final}} = (1 - \alpha)\, k_{\mathrm{DT}} + \alpha\, k_{\mathrm{root}}.$$

Section 3: Results

- We validated our method in UCF-Sports [3] and Highfive [4] datasets.
- **Node-VD (N) and Branch-VD (B) against Darwintrees (DT):** DT provided superior performance than N or B on UCF-Sports. On Highfive, DT demonstrated its complementarity with Root-VD.

Method	UCF [3]	Highfive [4] (mAP)		
		F#1	F#2	TOTAL
N	85.11	76.55	70.41	73.48
B	80.85	76.25	**72.53**	74.39
DT (N+B)	**91.49**	76.04	70.37	73.21
Root+DT	**91.49**	**79.24**	72.32	**75.78**

Table 1. Node-VD (N) Branch-VD (B) versus Darwintrees (DT) and DT combined with root (Root+DT) at kernel level.

- We also compared to other **SOTA methods**.

Method	Accuracy (%)
Ours (Root+DT)	**91.5**
Karaman et al. (2014)	90.8
Ma et al. (2015)	89.4
Wang et al. (2013)	85.2
Ma et al. (2013)	81.7
Raptis et al. (2012)	79.3

Table 2. Results on UCF-Sports dataset.

Method	mAP
Ours (Root+DT)	**75.8**
Wang et al. (2015)	69.4
Karaman et al. (2014)	65.4
Ma et al. (2015)	64.4
Gaidon et al. (2014)	62.4
Ma et al. (2013)	36.9
Patron-Pérez et al. (2012)	42.4

Table 3. Results on Highfive dataset.

Section 4: Conclusions

- A novel mid-level representation for action recognition on RGB videos.
- We modeled the evolution of features on both trajectory clusters and on the hierarchy defining those groupings.
- It is applicable to any local spatio-temporal feature representation.
- We demonstrated superior performance than other SOTA methods, especially for Highfive.

References

[1] H. Wang and C. Schmid. Action recognition with improved trajectories. In *CVPR*, pages 3551–3558, 2013.

[2] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, pages 5378–5387, 2015.

[3] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008.

[4] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. In *IEEE TPAMI*, 34(12):2441–2453, 2012.

[5] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. In *IJCV*, 107(3):219–238, 2014